Analyzing the Performance of LRU Caches under Non-Stationary Traffic Patterns

Mohamed Ahmed*, Stefano Traverso †, Paolo Giaccone†, Emilio Leonardi† and Saverio Niccolini*
* NEC Laboratories Europe, Heidelberg, Germany – {firstname.lastname}@neclab.eu
† Department of Electronics and Telecommunications, Politecnico di Torino, Torino, Italy – {lastname}@tlc.polito.it

***Abstract*—This work presents, to the best of our knowledge of the literature, the first analytic model to address the performance of an LRU (Least Recently Used) implementing cache under non-stationary traffic conditions, i.e., when the popularity of content evolves with time. We validate the accuracy of the model using Monte Carlo simulations. We show that the model is capable of accurately estimating the cache hit probability, when the popularity of content is non-stationary.**

We find that there exists a dependency between the performance of an LRU implementing cache and i) the lifetime of content in a system, ii) the volume of requests associated with it, iii) the distribution of content request volumes and iv) the shape of the popularity profile over time.

I. Introduction

Content caching is today a primitive network management operation, while the operation and performance of Content Delivery Networks (CDNs) is predicated on understanding on how users consume content.

This argument is given urgency by two main factors. First, the continued growth of Internet traffic, especially in the mobile context, increases the demand on limited network resources [1], [2]. Second, the dominance of multimedia traffic is today de-facto [3]. Recent studies [2] show that in the US, services offering real-time video and audio streaming occupy 62.5% and 54.7% of peak-period downstream traffic for fixed and mobile networks respectively. Similarly in Europe, real-time multimedia traffic accounts for 33.5-50% of peak-period downstream traffic in fixed networks, while globally, video traffic alone is project to account for 55% of all “consumer Internet traffic” by 2016 [1].

This traffic profile poses unique challenges to providing users with a reliable Quality of Service (QoS). For instance, Liu et. al. [3] report that 20% of users experience re-buffering when streaming contents, while 14% of users suffer significant delays before videos start to play. Furthermore, new networking paradigms such as ICN (Information-Centric Networking) are built on the implicit assumption of ubiquitous content caching [4], [5], such that small caches are co-located with routers in order to offset traffic latency.

Therefore, improving the effectiveness of content caching is paramount in aiding to address the problem of scaling the network while providing the necessary QoS to users. However, the vast majority of the studies on caching assume traffic patterns to be time-invariant, i.e., users browse through a large static catalogue of contents and make requests according to their different *static* popularity distributions - typically assumed to be Zipf.

In reality however, the popularity of contents varies over time and different contents exhibit a wide range of popularity evolution patterns [6], [7]. Contents tend to differentiate in i) when they start to attract user attention, ii) how much attention they attract and iii) how long they sustain the attraction. For instance, contents related sporting or geo-political events such as Olympic videos tend to enjoy a very short lifetime [2], reflecting users’ immediate interest in the topic. In contrast, some YouTube music videos keep on attracting user attention many months after being released [7]. Clearly, this observed behaviour is incompatible with time-invariant popularity models, and raises the need for more accurate tools that take into consideration the evolution of the popularity of contents over time.

This work presents the first steps in this direction. We extend the results of Che et al. [8] to take the time-variant popularity of contents explicitly into account, and present an approximated model of an LRU (Least Recently Used) caching under non-stationary traffic conditions. The accuracy of our model is validated against Monte Carlo simulations and shows that the non-stationarity of content popularity has a dramatic effect on caching performance. We find that even when cache sizes are small as in the context of ICNs, i.e., when cache dynamics change much faster than the popularity evolution, the performance of the cache is largely sensitive to popularity dynamics.

II. LRU under non-stationary conditions

In this section, we first present the assumptions that unpin our model (Sec. II-A). This is followed by the derivation of the cache hit probability under the non-stationary traffic scenario (Sec. II-B), with specific reference to scenarios with large (Sec. II-C) and small cache sizes (Sec. II-D).

A. A simple non-stationary traffic model

We start by assuming that contents are introduced into a catalogue (i.e. uploaded on some server) at random. For simplicity, this is taken to be according to a homogeneous Poisson process with rate γ. Furthermore, it is assumed that individual content popularity evolves (over time) according to some predetermined profile. Initially, it is assumed that all contents follow the same popularity profile, in Sec. IV we show that this assumption can be relaxed.

Let us now consider a generic content m, introduced into the catalogue at time τ_m, and whose popularity evolves over time according to:

$$\lambda_m(t) = V_m \lambda(t - \tau_m)$$

where $\lambda(t)$ represents the popularity profile and V_m a random mark (i.e. a random quantity) associated to content m. Popularity ($\lambda_m(t)$) in this context represents the instantaneous rate at which requests for a given content m arrive at the cache.

Requests are assumed to form an independent time-inhomogeneous Poisson processes and $\lambda(t)$ is taken to be an arbitrary function satisfying the following conditions: i) (positiveness) $\lambda(t) \geq 0 \ \forall t$ with $\lambda(0^+) > 0$, ii) (causality) $\lambda(t) = 0 \ \forall t < 0$, iii) (smoothness) $\lambda(t)$ continuous almost everywhere, iv) (integrability) $\int_0^\infty \lambda(t)dt = 1$. The average content lifetime can be computed as $L = \int_0^\infty t\lambda(t)dt$.

Observe that V_m represents the expected total number of requests (volume) induced by content m during its whole life in the system. More specifically, by construction, the total number of requests for content m is given by a Poisson distribution with an average of V_m. We assume that volumes of requests for different contents form an i.i.d. sequence of random variables distributed around some reference V. We denote $\phi_V(x) = \mathbb{E}[e^{xV}]$ to be the moment generating function of V and $\phi'_V(x)$ its first derivative.

Finally, the aggregate process of requests arriving to the cache is now by construction a Cox process [9] whose stochastic intensity is given by: $\Lambda(t) = \sum_m V_m \lambda(t - \tau_m)$.

B. Cache hit probability

In this section we extend the LRU approximation of Che et. al. [8] to our non-stationary traffic model in order to estimate the cache hit probability, i.e. the probability that a generic request finds the content in the cache.

Consider a cache capable of storing C distinct contents. Let $T_C(m)$ be the time needed for C distinct contents not including m to be requested by users. $T_C(m)$ therefore represents the *cache eviction time* for content m, i.e. after which point content m will be evicted from the cache. Che's approximation is centred on assuming that the cache eviction time ($T_C(m)$) is deterministic and independent from the selected content (m). This assumption has been given a theoretical justification in [10], where it is shown that, under a Zipf-like static popularity distribution, the coefficient of variation of the random variable representing $T_C(m)$ tends to vanish as the cache size grows. Furthermore, the dependence of the eviction time on m becomes negligible when the content catalogue is sufficiently large. The arguments given in [10] are easily extended to our non-stationary traffic model when γ and C are sufficiently large.

Returning to our non-stationary traffic model, we can now state our main result:

Theorem 1: Consider a cache of size C implementing LRU policy, operating under a non-stationary popularity model (as introduced in Sec. II-A) with total stochastic intensity: $\Lambda(t) = \sum_m V_m \lambda(t - \tau_m)$. Extending Che's approximation, the hit probability is given by:

$$p_{\text{hit}} = 1 - \int_0^\infty \lambda(\tau) \ \frac{\phi'_V \left(-\int_0^{T_C} \ \lambda(\tau - \theta)d\theta \right)}{\mathbb{E}[V]} d\tau \qquad (1)$$

where T_C is the solution to the equation:

$$C = \gamma \int_0^\infty 1 - \phi_V \ \left(- \int_0^{T_C} \lambda(\tau - \theta)d\theta \right) d\tau \qquad (2)$$

and γ is the rate at which new contents are introduced into the catalogue.

Proof: Proceeding along the same lines as for the stationary case (recalled in Appendix A), we consider a constant cache eviction time T_C. We can now start evaluating the probability of finding a given content in the cache at time t, conditional on the time it has been introduced into the catalogue (τ) and its request volume (V). This corresponds to the event in which one or more requests for the content occur in the time interval $[t - T_C, t]$. This probability can be evaluated as [1]: as:

$$p_{in}(t \mid \tau, V) = 1 - e^{-V \int_{t-T_C}^t \lambda(\theta - \tau)d\theta} \qquad (3)$$

Now, unconditioning with respect to V in (3), we obtain:

$$p_{in}(t \mid \tau) = \mathbb{E}_V \left[1 - e^{-V \int_{t-T_C}^t \lambda(\theta-\tau)d\theta} \right] = \\ 1 - \phi_V \left(- \int_{t-T_C}^t \lambda(\theta - \tau)d\theta \right)$$

To evaluate the probability of finding the given content in the cache at time t, we uncondition with respect to τ and obtain:

$$p_{in}(t) = \frac{1}{t} \int_0^t 1 - \phi_V \left(- \int_{t-T_C}^t \lambda(\theta - \tau)d\theta \right) d\tau \qquad (4)$$

This result exploits the elementary property of Poisson processes that, when a point falls within a specified interval of time, its distribution is uniform over the considered interval.

Now, as in the case of the stationary popularity scenario, for a sufficiently large t, the cache is completely filled with contents introduced to the catalogue before t and the number of contents in the cache is exactly equal to the size of the cache. We can therefore claim:

$$C = \sum_m [\mathbb{I}_{\{\text{content m in cache} \mid \tau_m \leq t\}} \mathbb{I}_{\tau_m \leq t}]$$

where τ_m is the time at which content m is introduced into the catalogue, and the sum extends over all the contents in the infinite content catalogue. Averaging both terms we obtain:

$$C = \sum_m \mathbb{E}[\mathbb{I}_{\{\text{content m in cache} \mid \tau_m \leq t\}} \mathbb{I}_{\tau_m \leq t}] = p_{in}(t) \sum_m \mathbb{E}[\mathbb{I}_{\tau_m \leq t}] \qquad (5)$$

[1] similarly to (19) obtained under stationary popularity

Since the average number of contents introduced to the catalogue at any time interval of size t is γ, by combining (4) with (5), we can evaluate the size of the cache C as:

$$C = \left(\sum_m \frac{\mathbb{E}[\mathbb{I}_{\tau_m < t}]}{t} \right) \int_0^t 1 - \phi_V \left(- \int_{t-T_C}^t \lambda(\theta - \tau) d\theta \right) d\tau = \gamma \int_0^t 1 - \phi_V \left(- \int_{t-T_C}^t \lambda(\theta - \tau) d\theta \right) d\tau \quad (6)$$

Equation (6) proves (2), and must be solved (numerically) to evaluate the eviction time (T_C) given a cache size of C.

Having defined T_C, we now return to evaluating the hit probability for a given content with the parameters (τ_0, V_0). By definition, the request for a given content at time t, generates a hit at the cache iff the content is located on the cache. Therefore, the probability of a hit is given by:

$$p_{hit}(t \mid \tau_0, V_0) = p_{in}(t \mid \tau_0, V_0)$$

Now, to uncondition $p_{hit}(t \mid \tau_0, V_0)$ with respect to V_0 and τ_0, we have to consider that the probability with which contents are requested by users is biased toward contents with higher instantaneous popularity. Let $N(V_0, \Delta V, \tau_0, \Delta\tau_0)$ be the average number of contents that have been generated during the interval $[\tau_0, \tau_0 + \Delta\tau_0)$ with request volume in $[V_0, V_0 + \Delta V_0)$. Now the probability that a request arrives for one of the contents defined above is

$$\frac{N(V_0, \Delta V, \tau_0, \Delta\tau_0)}{\gamma t} \times \frac{V_0 \lambda(t - \tau_0)}{\mathbb{E}[V]}$$

where the second term represents the instantaneous rate originated by every considered content. Thus, recalling (3) we have:

$$p_{\text{hit}}(t) = \mathbb{E}_{\tau,V} \left[\frac{V\lambda(t-\tau)}{\mathbb{E}[V]} p_{in}(t \mid \tau, V) \right] = \mathbb{E}_V \int_0^t \left[\frac{V\lambda(t-\tau)}{\mathbb{E}[V]} \left(1 - e^{-V \int_{t-T_C}^t \lambda(\theta-\tau)d\theta} \right) \right] d\tau = \int_0^t \lambda(t-\tau)\mathbb{E}_V \left(\frac{V}{\mathbb{E}[V]} - \frac{Ve^{-V \int_{t-T_C}^t \lambda(\theta-\tau)d\theta}}{\mathbb{E}[V]} \right) d\tau = \int_0^t \lambda(t-\tau) \left(1 - \frac{\phi_V' \left(-\int_{t-T_C}^t \lambda(\theta-\tau)d\theta \right)}{\mathbb{E}[V]} \right) d\tau$$

By substituting $\alpha = t - \tau$ and $\beta = t - \theta$:

$$p_{\text{hit}}(t) = \int_0^t \lambda(\alpha) \left(1 - \frac{\phi_V' \left(-\int_0^{T_C} \lambda(\alpha - \beta) d\beta \right)}{\mathbb{E}[V]} \right) d\alpha \quad (7)$$

Thanks to the integrability property of $\lambda(t)$, (1) is obtained by letting $t \to \infty$ in (7). ■

The following corollary sheds some light on the relation between C and T_C:

Corollary 1: The variables C and T_C satisfy:

$$C \leq \gamma\mathbb{E}[V] \int_0^\infty \int_0^{T_C} \lambda(\tau - \theta) d\theta d\tau = \gamma\mathbb{E}[V]T_C \quad (8)$$

and

$$C \geq \gamma\mathbb{E}[V]T_C - \gamma\frac{\mathbb{E}[V^2]}{2} \int_0^\infty \left(\int_0^{T_C} \lambda(\tau - \theta) d\theta \right)^2 d\tau \quad (9)$$

Proof: The inequality in (8) is derived for (2) by exploiting the inequality $1 + x \leq e^x$. In particular we exploit the previous inequality to lower bound $\phi_V(-\int_0^{T_C} \lambda(\tau - \theta)d\theta) = \mathbb{E}[e^{-\int_0^{T_C} \lambda(\tau-\theta)d\theta)}]$ with $1 - \mathbb{E}[V] \int_0^{T_C} \lambda(\tau - \theta)d\theta$ inside the integral appearing in (2).

Similarly, the inequality in (9) is obtained by exploiting $1 - x + \frac{x^2}{2} \geq e^{-x}$ for every $x \geq 0$. ■

Note that the upper bound in (8) has a simple meaning; assuming that all the requests are referring to different contents, the cache size is bounded by the overall number of requests $\gamma E[V]$ during the time interval T_C. In Sec. II-D, we will show that this bound is a good approximation for C when cache size is small.

To gather more insights on the impact of different parameters on p_{hit}, we now derive a simplified expression for the two extreme scenario regimes of *large* cache and *small* cache sizes.

C. Large-cache regime

A closed form expression for the asymptotic hit probability ($p_{\text{hit},\infty}$) when the cache $C \to \infty$ can be derived from (1) by making $T_C \to \infty$.

Corollary 2: For large cache sizes,

$$p_{\text{hit},\infty} = 1 - \frac{1 - \phi_V(-1)}{\mathbb{E}[V]} = 1 - \frac{1}{\mathbb{E}[V]} + \frac{\mathbb{E}[e^{-V}]}{\mathbb{E}[V]} \quad (10)$$

Proof: Consider the limit as $T_C \to \infty$ in the integral within (1); it holds that:

$$\int_0^\infty \lambda(\tau)\phi_V' \left(-\int_0^\infty \lambda(\tau - \theta)d\theta \right) d\tau = \int_0^\infty \lambda(\tau)\phi_V' \left(-\int_0^\tau \lambda(\alpha)d\alpha \right) d\tau \quad (11)$$

Now if we define $\Lambda(\tau) = \int_0^\tau \lambda(\alpha)d\alpha$ (by construction, $\Lambda(\alpha)$ is also the primitive of $\lambda(\alpha)$) and $\beta = \Lambda(\tau)$, by substituting β into (11), we obtain:

$$\int_0^\infty \lambda(\tau)\phi_V'(-\Lambda(\tau))d\tau = \int_0^1 \phi_V'(-\beta)d\beta = \phi_V(0) - \phi_V(-1) = 1 - \phi_V(-1) \quad (12)$$

Finally, (10) is obtained by using (12) within (1). ■

Observe that (10) depends heavily on the distribution of the content request volumes (V), and is completely independent of the temporal profile of the popularity ($\lambda(t)$). This is expected, when we consider that as C and T_C grow large, contents are never evicted from the cache. In effect, only the first request for every content will lead to a cache miss, independently of the arrival request pattern.

The expression (10) is exact, since for $C \to \infty$ it can be easily proved that $T_C(m) \to \infty$ w.p.1. This is obtained exploiting by the following properties: i) as $C \to \infty$ the conditional hit probability for contents originating $R \geq 1$ requests tends to $p_{\text{hit}}(R) = 1 - 1/R$, and ii) the probability of observing at least one request for content m is $\Pr(R \geq 1) = 1 - e^{V_m}$.

The value of C (and consequently T_C) for which p_{hit} approaches $p_{\text{hit},\infty}$, instead heavily depends the popularity profile $\lambda(t)$. Indeed it is possible to derive a bound on the difference of the hit probability from $p_{\text{hit},\infty}$:

Corollary 3:

$$p_{\text{hit},\infty} - p_{\text{hit},T_C} \leq \int_{T_C}^{\infty} \lambda(\tau) d\tau$$

Proof: Starting from (1), we obtain:

$$\begin{aligned}
p_{\text{hit}} = 1 - \int_0^{\infty} \lambda(\tau) \frac{\phi'_V\left(-\int_0^{T_C} \lambda(\tau-\theta)d\theta\right)}{\mathbb{E}[V]} d\tau = \\
\int_0^{\infty} \lambda(\tau) \left[1 - \frac{\phi'_V\left(-\int_0^{T_C} \lambda(\tau-\theta)d\theta\right)}{\mathbb{E}[V]}\right] d\tau = \\
\int_0^{T_C} \lambda(\tau) \left[1 - \frac{\phi'_V\ \left(-\int_0^{\tau} \lambda(\alpha)d\alpha\right)}{\mathbb{E}[V]}\right] d\tau + \\
\int_{T_C}^{\infty} \lambda(\tau) \left[1 - \frac{\phi'_V\left(-\int_0^{T_C} \lambda(\tau-\theta)d\theta\right)}{\mathbb{E}[V]}\right] d\tau \quad (13)
\end{aligned}$$

where we have operated the change of variable $\alpha = \tau - \theta$. By observing that $\phi'(x) = \mathbb{E}[Ve^{xV}] \leq \mathbb{E}[V]$ for any $x \leq 0$, it is possible to upper bound the right-most term of (13) as follows:

$$\int_{T_C}^{\infty} \lambda(\tau) \left[1 - \frac{\phi'_V\left(-\int_0^{T_C}\ \lambda(\tau-\theta)d\theta\right)}{\mathbb{E}[V]}\right] d\tau \leq \int_{T_C}^{\infty} \lambda(\tau)d\tau \quad (14)$$

Thanks to (12) it is also possible to upper bound the left-most term:

$$\begin{aligned}
\int_0^{T_C} \lambda(\tau) \left[1 - \frac{\phi'_V\left(-\int_0^{\tau} \lambda(\theta)d\theta\right)}{\mathbb{E}[V]}\right] d\tau \leq \\
\int_0^{\infty} \lambda(\tau) \left[1 - \frac{\phi'_V\left(-\int_0^{\tau} \lambda(\theta)d\theta\right)}{\mathbb{E}[V]}\right] d\tau = \\
1 - \int_0^{\infty} \lambda(\tau) \frac{\phi'_V(-\Lambda(\tau))}{\mathbb{E}[V]} d\tau = 1 - \frac{1 - \phi_V(-1)}{\mathbb{E}[V]} \quad (15)
\end{aligned}$$

which corresponds to $p_{\text{hit},\infty}$. By combining (14) with (15), we get the assert. ■

D. Small-cache regime

Under this regime, we get the following hit probability:

Corollary 4: For very small cache sizes, we can approximate the hit probability as:

$$p_{\text{hit}} \approx \frac{\mathbb{E}[V^2]}{\mathbb{E}[V]} T_C \int_0^{\infty} \lambda^2(\tau) d\tau \quad (16)$$

Proof: The expression in (16) is obtained from (1) by assuming $\int_0^{T_C} \lambda(\tau - \theta)d\theta \approx \lambda(\tau)T_C \ll 1$, and approximating $\phi'_V(x) = \mathbb{E}[Ve^{xV}] \approx \mathbb{E}[V] + x\mathbb{E}[V^2]$ for small values of x. ■

Profile	$\lambda(t)$	$\int_0^{\infty} \lambda^2(\tau)d\tau$
Exponential	$\frac{1}{L}e^{-t/L}$ for $t \geq 0$	$\frac{1}{2L}$
Power law ($\zeta > 1$)	$\frac{\zeta-1}{L}\left(\frac{t}{L}+1\right)^{-\zeta}$ for $t \geq 0$	$\frac{(\zeta-1)^2}{L(2\zeta-1)}$
Uniform	$\frac{1}{2L}$ for $t \in [0, 2L]$	$\frac{1}{2L}$
Triangular	$\begin{cases} \frac{t}{L^2} & \text{for } t \in [0, L] \\ \frac{2L-t}{L^2} & \text{for } t \in [L, 2L] \end{cases}$	$\frac{2}{3L}$

TABLE I

EXAMPLES OF POPULARITY PROFILES $\lambda(t)$. FOR ALL PROFILES THE AVERAGE CONTENT LIFETIME IS SET EQUAL TO L. OBSERVE THAT $\int_0^{\infty} \lambda^2(\tau)d\tau$ IS THE KEY PARAMETER APPEARING IN (18).

Furthermore, we can improve the results in Corollary 1 to better approximate the relation between C and T_C as follows:

$$C \approx \gamma \mathbb{E}[V] T_C \quad (17)$$

Following the same reasoning as the proof for (8), observe that, for small values of T_C, $\int_0^{T_C} \lambda(\tau - \theta)d\theta \ll 1$. Now $\phi_V(-\int_0^{T_C} \lambda(\tau - \theta)d\theta)$ can be approximated with $1 - \mathbb{E}[V] \int_0^{T_C} \lambda(\tau - \theta)d\theta$ and the desired relation is obtained.

Using (17), we can now rewrite (16) as follows:

$$p_{\text{hit}} \approx \frac{\mathbb{E}[V^2]}{\mathbb{E}^2[V]} \frac{C}{\gamma} \int_0^{\infty} \lambda^2(\tau) d\tau \quad (18)$$

The expression given in (18) enlightens us to the potentially large effect the popularity profile of content has on the effectiveness of caching, when cache sizes are small. For illustration, if we consider the popularity profiles given in Table I, the hit probability is always inversely proportional to the average content lifetime (L). But, by comparing the third column, it is clear that the actual value depends strongly on the shape of the profile.

III. NUMERICAL VALIDATION

In this section we present: i) the results of applying the approximation of LRU under non-stationary traffic (see Sec.II-A), as given by Theorem 1; ii) the validation of the predictions of the model through Monte Carlo simulations of a single cache.

The results presented in this section relate the size of a cache (C) to the hit probability (p_{hit}) and look at relation between these two variables, when varying: i) the average content lifetime (L), ii) the average content request volume ($\mathbb{E}[V]$), iii) the distribution of content request volumes, iv) the shape of the popularity time profile.

For all the results that follow, we set the content arrival rate (γ) to 10k contents per day and the content request volume (V) is assumed to be distributed according to a Pareto distribution: $f_V(v) = \beta V_{\min}^{\beta}/v^{1+\beta}$ for $v \geq V_{\min}$. The choice of a Pareto distribution is justified by two factors. First, several recent measurement studies have shown that the Zipf law is a very good approximation of the empirical distribution of long term content (video) request volumes [10], [11]. Second, a Zipf-like distribution of content request volumes with parameter

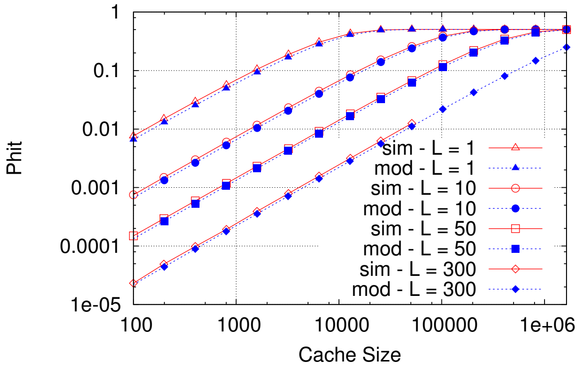

Fig. 1. Cache hit probability under exponential popularity profile, for different values of the average content lifetime L (expressed in days).

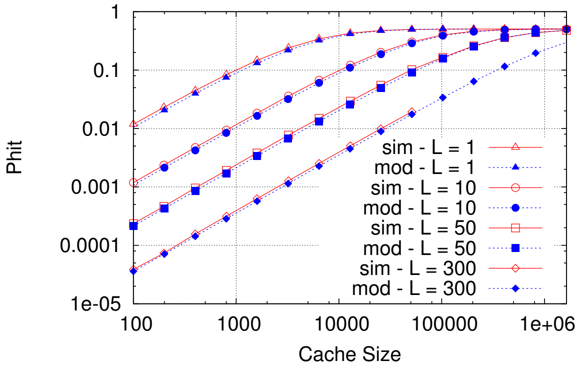

Fig. 2. Cache hit probability under power law popularity profile with $\zeta = 3$, for different values of the average content lifetime L (expressed in days).

$\alpha = 1/(\beta - 1)$ is obtained when a large number of individual content request volumes are independently generated according to Pareto distribution.

With regard to the different popularity profiles that content may display, we consider exponential and power law profiles as given in Table I. Finally, the parameter ζ is used to model the different time-dependent popularity profiles (popularity shapes).

Figs. 1 and 2 report the hit probability for different values of content lifetime, with respect to using; i) an exponential (see Fig. 1) and ii) power law with parameter $\zeta = 3$ (see Fig. 2) popularity profile. In both cases, we set $V_{\min} = 1$ and $\beta = 3$, as a consequence, we obtain $\mathbb{E}[V] = 1.5$ requests per content.

The first point to observe is that the model estimates agree strongly with the simulation results for the hit probability in all the cases. Second, as predicted by the model, the average content lifetime (L) deeply impacts the cache performance. Indeed, for a given cache size (C), a given content's hit probability increases significantly as its lifetime is reduced. In particular, for moderate cache sizes, the hit probability is roughly inverse proportional to the content lifetime, as predicted by Corollary 4.

Fig. 3 reports the cache hit probability for different values of the average content request volume $\mathbb{E}[V]$. All plots refer to the same value of $\beta = 3$ and different values of $V_{\min} = \mathbb{E}[V]\frac{\beta-1}{\beta}$.

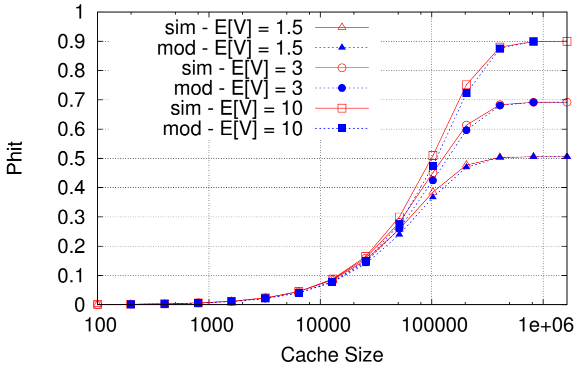

Fig. 3. Cache hit probability under exponential popularity profile with $L = 10$ days for different values of the average content volume $E[V]$.

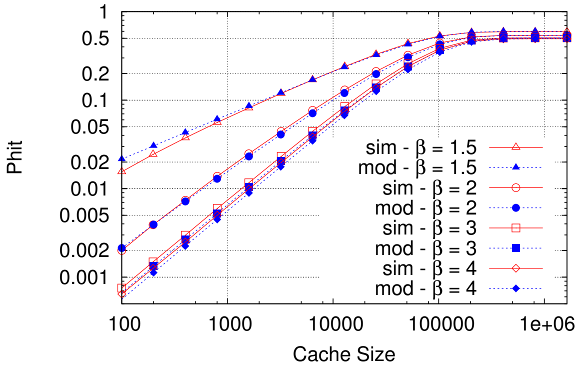

Fig. 4. Cache hit probability under exponential popularity profile for different request volume distributions and average content request volume $E[V] = 1.5$.

The results here reveal that the volume of hits accumulated by content has impact exclusively for large cache sizes, i.e. when $C \geq$ 200k objects, and as predicted by (10), the cache hit probability increases with the volume of content requests. For moderate cache sizes ($C \leq$ 10k), the effect of the hits becomes negligible as predicted by (18). Indeed, all curves correspond to the same value of $\frac{\mathbb{E}[V^2]}{\mathbb{E}^2[V]}$.

Fig. 4 reports the cache hit probability for different values of the parameter β, associated to the distribution of content request volumes. From the figure we see that the shape of requests volumes has a significant impact on the cache hit probability. As expected, by decreasing β (i,e., increasing the correspondent parameter α for the associated Zipf law), the cache performance is improved. In particular we observe that the caching performance become much more sensitive to β as β gets smaller than 2 (i.e., as the corresponding Zipf parameter α increases above 1). However, the impact of β (i.e., α) on caching performance does not appear in our scenario to be as strong as it does in the classical stationary case, where a sort of "phase transition" is observed as α crosses the value 1 [10].

Finally, Fig. 5 reports the cache hit probability for different content popularity profiles (i.e, varying ζ) - while keeping the average lifetime as constant at $L = 10$ days. From the figure, we see that the content popularity profile appears to have a moderate impact on the caching performance (for small

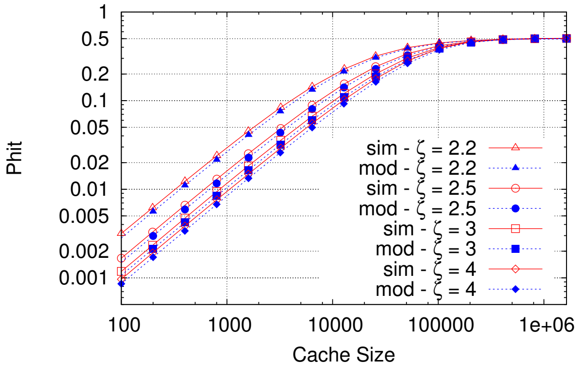

Fig. 5. Cache hit probability under power law popularity profile for different values of ζ and the same average content lifetime $L = 10$ days.

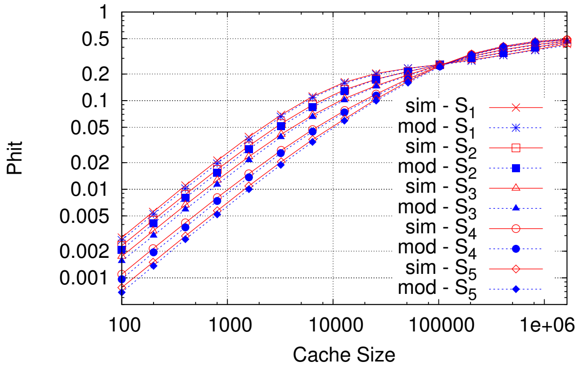

Fig. 6. Cache hit probability under exponential popularity profile for different classes configurations.

caches) as long as the average content lifetime L is kept constant. For the extreme case where the size cache $C = 100$, p_{hit} varies from 0.0032 and 0.001 as ζ is decreased from 4 to 2.2 (see the simulation curves).

IV. Extension to a Multi-class scenario

In this section, we consider a more realistic scenario in which contents can be partitioned into K classes, such that each class is associated with a different popularity profile $\lambda_k(t)$ and a different request volume distribution V_k, for $1 \leq k \leq K$.

This generalisation of our traffic model is needed in order to capture the variability of the popularity profiles exhibited by real contents. Indeed the popularity profile depends heavily on the nature of contents, for example, the popularity evolution of music videos is typically significantly different to videos containing sport highlights. However, recent experimental studies [7] have shown that the popularity evolution of different contents can be clustered to relatively few groups exhibiting similar temporal popularity profiles.

We can formalise the multi-class scenario by assuming that every generated content (m) is associated with a random mark W_m taking values in $\{1, \ldots, K\}$, such that the mark specifies the class the content belongs to. Assuming $\{W_m\}$ to be i.i.d. random variables, the total stochastic intensity at time t of the request process is given by:

$$\Lambda(t) = \sum_m V_m \lambda_{W_m}(t - \tau_m) = \sum_m V_m \lambda_{W_m}(t - \tau_m) \mathbb{I}_{\{\tau_m \leq t\}}$$

Under this assumption, we can now state the following:

Theorem 2: Consider a cache of size C implementing LRU policy, operating under a multi-class non-stationary popularity model, with total stochastic intensity: $\Lambda(t) = \sum_m V_m \lambda_{W_m}(t - \tau_m)$. Extending Che's approximation, the hit probability is given by:

$$p_{\text{hit}} = 1 - \sum_{k=1}^{K} \Pr\{W_1 = k\} \int_0^\infty \lambda_k(\tau) \frac{\phi'_{V_k}\left(-\int_0^{T_c} \lambda_k(\tau - \theta) d\theta\right)}{\mathbb{E}[V_k]} d\tau$$

where T_C is the solution to the equation:

$$C = \gamma \int_0^\infty \left[1 - \sum_1^K \Pr\{W_1 = k\} \phi_{V_k}\left(-\int_0^{T_C} \lambda_k(\tau - \theta) d\theta\right)\right] d\tau$$

Popularity class	L [days]	S_1	S_2	S_3	S_4	S_5
1	1	40%	30%	20%	10%	5%
2	10	10%	20%	30%	40%	45%
3	50	10%	20%	30%	40%	45%
4	300	40%	30%	20%	10%	5%

TABLE II
Average content lifetime L and fractions of contents in each class for setups from S_1 to S_5.

The proof for Theorem 2 (not reported here for the sake of brevity) follows the same lines of Theorem 1.

In order to evaluate the multi-class scenario, we partition contents into $K = 4$ classes, each characterised by a different popularity profile and lifetime. For simplicity, here, we only considered 5 different configurations as specified in Table II.

Fig. 6 reports the cache hit probability for each setup. First, from the figure, we see that the model predictions of the cache hit probability for different cache sizes align accurately with the results of the simulation. Second, Fig. 6 shows that the heterogeneity of contents weakly impacts the cache performance, and that the cache hit probability increases when the average content lifetime (L) of each setup decreases, i.e. when the number of contents with fast popularity dynamics (belonging to class 1) is large with respect to the rest. In fact, as already seen in Fig. 1, contents attracting users' attention for very limited periods of time ($L = 1$ days) are largely responsible for improving the cache performance when adopting the LRU strategy.

V. Conclusions

This work has proposed (and validated with a Monte Carlo simulation) a simple but highly accurate approximated model for an LRU (Least Recently Used) cache under non-stationary traffic conditions. The proposed model is flexible and can be easily extended to consider complex and realistic traffic scenarios. It has the advantage of being computationally cheap when compared with the Monte Carlo simulations - especially when cache sizes are large.

Our results show that caching performance is deeply impacted by the dynamics resulting from the popularity of content. In particular, we find that when cache sizes are small as in the case anticipated for ICN networks, the content hit probability is largely sensitive to the popularity profile of contents.

References

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2011-2016," Cisco, Tech. Rep., 2012. [Online]. Available: http://goo.gl/7AOgt

[2] Sandvine, "Global Internet Phenomena Report," Sandvine, Tech. Rep., 2012. [Online]. Available: http://goo.gl/djUqh

[3] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang, "A case for a coordinated internet video control plane," in *SIGCOMM*, 2012.

[4] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache "Less for More" in Information-Centric Networks," in *Networking*, 2012.

[5] D. Rossi and G. Rossini, "On sizing CCN content stores by exploiting topological information," in *NOMEN*, 2012.

[6] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *WSDM*, 2011.

[7] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: Predicting the evolution of popularity in user generated content," in *WSDM*, 2013.

[8] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305 – 1314, Sep. 2002.

[9] D. Cox and V. Isham, *Point processes*. Chapman & Hall/CRC, 1980, vol. 12.

[10] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," *CoRR*, vol. abs/1202.3974, 2012.

[11] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.

Appendix A
LRU under stationary popularity

First we briefly resume Che's approximation of LRU in a classical traffic scenario in which every content m, in a finite catalog of size M, presents a time-invariant popularity profile. More in details, we assume that requests for content m arrive at the cache according to a time homogeneous Poisson process with intensity λ_m. Let $\Lambda = \sum_m \lambda_m$ be the resulting total arrival intensity of content requests at the cache.

Now, thanks to Che's approximation, discussed in Sec. II-B, a content m is present in the cache at time t, if and only if a time less than T_c has passed since the last request for content m, i.e., if at least a request for such content has arrived in the interval $(t - T_c, t]$. Since requests arrivals are Poisson, the probability $p_h(m)$ that at at least one request has arrived in the interval $(t - T_c, t]$ is given by: $p_h(m) = 1 - e^{-\lambda_m T_c}$. Observe that $p_h(m)$ represents, by construction, also, the hit probability for content m, as immediate consequence of PASTA property of arrivals.

Considering a cache of size C, by construction: $C = \sum_m \mathbb{I}_{\{m \text{ in cache}\}}$. When averaging both sides, we obtain:

$$C = \sum_m \mathbb{E}[\mathbb{I}_{\{m \text{ in cache}\}}] = \sum_m p_h(m) = \sum_m (1 - e^{-\lambda_m T_c}).$$

By solving the above relationship, we obtain T_c, and then the average hit probability on the cache as:

$$p_{\text{hit}} = \sum_m p_m p_h(m) \tag{19}$$

A. Numerical evaluation

We test the Che's approximation in the stationary popularity case. We set a catalog size equal to $M = 10^7$ contents. Fig. 7 shows the impact of the Zipf's exponent α on the caching performance.

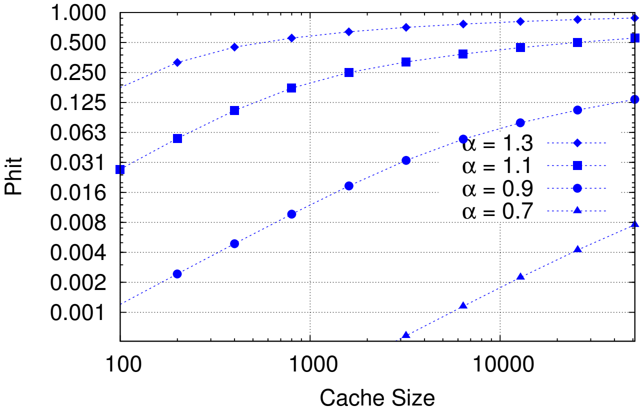

Fig. 7. Hit probability p_{hit} vs. cache size for different content popularity distributions

As already well known, the popularity distribution shape has a disruptive impact on caching performance.

For α sufficiently larger than 1, the popularity distribution is sufficiently skewed, so that the contribution of the few top popular contents to the total traffic is significant. Indeed, observe that that $H(M) = \Theta(1)$ when M grows large, i.e. the aggregate contribution of tail contents is marginal.

For $\alpha < 1$, instead, only caching a significant portion of the huge catalog we can achieve significant cache hit probability. In this case $H(M)$ diverges as $M \to \infty$ showing that the aggregate contribution of tail contents is dominant.